

SHORT COMMUNICATION

arakno - An R package for effective spider nomenclature, distribution and trait data retrieval from online resources

Pedro Cardoso¹ and **Stano Pekár**²: ¹Laboratory for Integrative Biodiversity Research - LIBRe, Finnish Museum of Natural History Luomus, PO17 (Pohjoinen Rautatiekatu 13), 00014, University of Helsinki, Finland. E-mail: pedro.cardoso@helsinki.fi; ²Department of Botany and Zoology, Faculty of Science, Masaryk University, Kotlářská 2, 611 37 Brno, Czechia.

Abstract. Online open databases are increasing in number, usefulness, and ease of use. There are currently two main global databases for spiders, the World Spider Catalogue (WSC) and the World Spider Trait (WST) database. Both are regularly used by thousands of researchers. Computational tools that allow effective processing of large data are now part of the workflow of any researcher and R is becoming a *de facto* standard for data manipulation, analysis, and presentation. Here we present an R package, *arakno*, that allows interface with the two databases. Implemented tools include checking species names against nomenclature of the WSC, obtaining and mapping data on distribution of species from both the WST and the Global Biodiversity Information Facility (GBIF), and downloading trait data from the WST. A set of tools are also provided to prepare data for further statistical analysis.

Keywords: Araneae, biological traits, functional diversity, geographical range, taxonomy.

<https://doi.org/10.1636/JoA-S-21-024>

Online open databases are increasing in number, usefulness, and ease of use. In this respect, spider researchers have always been at the forefront, with the World Spider Catalogue (2021; hereafter WSC) serving as a continuously updated reference on the nomenclature of the group. The WSC provides taxonomic information including the most updated nomenclature, previous nomenclature, synonyms or misidentifications in past taxonomic works. The WSC also allows registered users to freely and immediately download any publication of taxonomic relevance, a privilege that almost no other researchers have. Including all spider species ever described, currently close to 50,000, the WSC is and will continue to be extensively used for multiple purposes within and outside research, from checking current nomenclature in faunistic works, to serving as the taxonomic basis for major conservation efforts worldwide (e.g., Seppala et al. 2018). Its data are free to use and already in 2015 logged a daily average of 600 hits and 400 downloads (Nentwig et al. 2015).

The World Spider Trait database (Pekár et al. 2021; hereafter WST) is a recent addition to the arsenal of virtual infrastructure available to arachnology. It intends to compile curated data on spider traits (such as morphological, ecological, physiological, or behavioural) and this way both serve as a repository for data used in many publications and promote the use of trait data for eco-evolutionary analyses (Lowe et al. 2020). Its data are provided under a CC BY 4.0 license, and given its extremely recent availability, no relevant statistics are available yet.

Both databases store huge amounts of data that can be quickly accessed via online applications (see <https://wsc.nmbe.ch/> and <https://spidertraits.sci.muni.cz/>, respectively). These applications provide simple tools for data retrieval if the amount of data to be retrieved is relatively small. However, as the number of species of interest increases, as in, for example, community ecology studies, the use of online tools becomes ineffective.

Here we present the R package *arakno* (ARACHnid KNowledge Online; Cardoso 2021) available from CRAN (online at <https://CRAN.R-project.org/package=arakno>). It includes a suite of functions within the R environment (R Core Team 2021) allowing download of data from both the WSC and WST, checking valid

species names, listing or mapping species distributions, and preparing trait data for further analysis. The R language is becoming a *de facto* standard for data manipulation, analysis, and results presentation in ecology and evolution, and we expect these functions will be of use for many arachnologists, among other researchers. Specifically, taxonomists or community ecologists can use them to assemble lists of valid species names with authorities and family membership or to produce maps of current known distribution. Evolutionary biologists, community ecologists, macroecologists, among many others, can download trait data to test predefined hypotheses.

The package *arakno* contains the following functions (arguments):

wsc() – Downloads the current, daily updated, list of valid spider names with their distribution data (list of countries) from WSC and returns a *data.frame* named *wscdata* to the global R environment.

checknames(tax, full = FALSE, order = FALSE) – Compares binomial species, genus and family names provided in a vector named *tax* with the nomenclature from WSC and, in case mismatches are detected, returns a four column array with original names (Species), possible matches (Best match; Alternative match(es)) and a Note indicating whether a nomenclature change, junior synonym or misspelling was detected. Misspellings in names are matched using fuzzy logic (minimum Levenshtein edit distance, representing the minimum number of character deletions, additions or substitutions needed to convert one word into another). Subspecies names are not eligible. If *full = TRUE* all species names are returned, including valid names. As in other functions of this package, the results can be ordered alphabetically or follow the order given in *tax* (the default).

authors(tax, order = FALSE) – Returns authority and year for binomial species names provided in *tax* according to the WSC. If genera or families are used, data on all their species are retrieved.

distribution(tax, order = FALSE) – Returns distributions (countries or geographic areas) of species given in *tax* from the WSC. If genera or families are used, data on all their species are retrieved.

lsid(tax, order = FALSE) – Returns the *LSID* (Life Science Identifier, a unique code for each species) of species given in *tax* from the

WSC. If genera or families are used, data on all their species are retrieved.

species(tax, order = FALSE) – Returns all species currently belonging to a family or genus given in *tax* according to the WSC.

taxonomy(tax, check = FALSE, aut = FALSE, id = FALSE, order = FALSE) – Returns the classification at sub/infraclass, family, and genus level for species in *tax* according to the WSC. If genera or families are used, data on all their species are retrieved. Optionally, if *check = TRUE* the function will check the names (similar to function *checknames*); if *aut = TRUE* the function will provide authorities (similar to function *authors*); and if *id = TRUE* the function will give the LSIDs from the WSC (similar to function *lsid*).

traits(tax, trait = NULL, sex = NULL, life = NULL, country = NULL, habitat = NULL, user = “.”, key = “.”, order = FALSE) – Downloads all available trait data from the WST for all taxa given in *tax*. If genera or families are used, data on all their species are retrieved. Eligible values for *trait*, *sex*, *life*, *country*, and *habitat* are available online at <https://spidertraits.sci.muni.cz/traits>. For traits, the abbreviation is used. To access embargoed data, a *user* name and API (Application Programming Interface) *key* from the WST must be provided.

records(tax, order = FALSE) – Returns coordinates and sources of occurrences for the taxa given in *tax* from both the WST and the Global Biodiversity Information Facility (GBIF; GBIF.org 2021). GBIF allows easy access to distribution data from multiple institutions worldwide, with the licensing being determined for each dataset. If genera or families are used, data on all their species are retrieved. The user should be aware that coordinate data returned are incomplete, often suffering from the Wallacean shortfall, i.e., incomplete geographical coverage (Lomolino 2004; Cardoso et al. 2011).

map(tax, countries = TRUE, records = TRUE, hires = FALSE, zoom = FALSE, order = FALSE) – Returns a map of distribution for the species given in *tax* using a list of *countries* and regions according to the WSC and/or *records* with coordinates available from the WST and GBIF. If genera or families are used, data on all their species are retrieved. Maps can be of high resolution (*hires = TRUE*) and the area zoomed (*zoom = TRUE*) to the area covered by the *records* available for each species.

Use of all functions is demonstrated in Supplementary file Appendix 1 (online at <https://doi.org/10.1636/JoA-S-21-024.s1>). Additional functions are planned to be implemented in future releases of the package *arakno*. Specifically:

- 1) Connection with other arachnid databases, such as the World Catalog of Opiliones (Kury et al. 2020). This will only be possible when digital identifiers and APIs are available, so we would like to appeal to authors of the catalogs to implement these requirements.
- 2) Connection with the International Union for the Conservation of Nature Red List (IUCN 2021) once arachnid coverage is comprehensive.
- 3) Seamless integration of *arakno* with other packages developed by the same author for ecology, biogeography and conservation purposes (Cardoso et al. 2015; Cardoso 2017), for an improved workflow.

The future development and utility of functions within the package are dependent on maximizing species and trait data coverage, data standardization and enable connectivity of databases according to the FAIR principles - Findability, Accessibility, Interoperability, and Reusability (Wilkinson et al. 2016). By following these principles, it is possible to easily connect *arakno* with multiple databases independently of their specific goal or format, this way maximizing data usage. We therefore strongly encourage:

- 1) Arachnid database managers to follow the FAIR principles so that data can be fully explored for multiple purposes.
- 2) Arachnid experts to contribute to the multiple databases (e.g., Pekár et al. 2021) to both guarantee archiving and re-use of data and contribute to global efforts on species knowledge and conservation.

The R package *arakno* tries to fill a gap in facilitating access to online open databases using robust, flexible, and expandable functions. It is open source and any contributions and suggestions to its continuous development are most welcome.

ACKNOWLEDGMENTS

We thank all the developers and data providers of the WSC and WST. Adam Kučera is thanked for development of the WST API. Michael Rix for final edits. The Pentti Tuomikoski Grant from the Finnish Museum of Natural History provided financial support to the development of *arakno* and the WST.

SUPPLEMENTAL MATERIALS

Appendix 1.—Three examples from the *arakno* package, online at <https://doi.org/10.1636/JoA-S-21-024.s1>

LITERATURE CITED

- Cardoso P. 2017. Red – an R package to facilitate species red list assessments according to the IUCN criteria. *Biodiversity Data Journal* 5:e20530. doi: 10.3897/BDJ.5.e20530
- Cardoso P. 2021. *arakno* – ARACHnid KNowledge Online. Version 1.1.0. Online at <https://CRAN.R-project.org/package=arakno>, accessed on 2021.06.06.
- Cardoso P, Erwin TL, Borges PAV, New TR. 2011. The seven impediments in invertebrate conservation and how to overcome them. *Biological Conservation* 144:2647–2655. doi: 10.1016/j.biocon.2011.07.024
- Cardoso P, Rigal F, Carvalho JC. 2015. BAT - Biodiversity Assessment Tools, an R package for the measurement and estimation of alpha and beta taxon, phylogenetic and functional diversity. *Methods in Ecology and Evolution* 6:232–236. doi: 10.1111/2041-210X.12310
- GBIF.org. 2021. GBIF Home Page. Online at <https://www.gbif.org>, accessed on 2021.04.06.
- IUCN. 2021. The IUCN Red List of Threatened Species. Version 2021–1. Online at <https://www.iucnredlist.org/>, accessed on 2021.04.06.
- Kury AB, Mendes AC, Cardoso L, Kury MS, Granado AA, Giribet G. 2020. World Catalogue of Opiliones, WCO-Lite. Version 1.2.1. Online at <https://wcolite.com/>, accessed on 2021.04.06.
- Lomolino MV. 2004. Conservation biogeography. Pp. 293–296. *In* Frontiers of Biogeography: New Directions in the Geography of Nature. (Lomolino MV, Heaney LR. (eds.)), Sinauer Associates, Sunderland, Massachusetts.
- Lowe E, Wolff JO, Aceves-Aparicio A, Birkhofer K, Branco VV, Cardoso P. et al. 2020. Towards establishment of a centralized spider traits database. *Journal of Arachnology* 48:103–109. doi: 10.1636/0161-8202-48.2.103.
- Nentwig W, Gloor, D, Kropf C. 2015 Spider taxonomists catch data on web. *Nature* 528: 479. doi: 10.1038/528479a
- Pekár S, Wolff J, Černecká Ľ, Birkhofer K, Mammola S, Lowe EC, et al. 2021. The World Spider Trait database (WST): a centralised global open repository for curated data on spider traits. *Database* 2021 <https://doi.org/10.1093/database/baab064>
- R Core Team. 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna.

- Online at <http://www.r-project.org/index.html>, accessed on 2021.04.06.
- Seppälä S, Henriques S, Draney ML, Foord S, Gibbons AT, Gomez LA, et al. 2018. Species conservation profiles of a random sample of world spiders IV: Scytodidae to Zoropsidae. *Biodiversity Data Journal* 6: e3842. doi: 10.3897/BDJ.6.e30842
- Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3:160018. doi: 10.1038/sdata.2016.18
- World Spider Catalog. 2021. World Spider Catalog. Version 22.0 Natural History Museum, Bern. Online at <http://wsc.nmbe.ch>, accessed on 2021.04.06. doi: 10.24436/2
- Manuscript received 19 April 2021, revised 22 June 2021, accepted 23 June 2021.*